

Statistiques descriptives de mono variable

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

3 février 2026

Plan

Description mono variable

- La démarche descriptive

- Tableau de données

- Résumé (statistique) des données

 - Tendance centrale

 - Dispersion

 - Autres moments

 - Résumé robuste

- Résumé graphique des données

 - Cas des variables qualitatives

 - Boite à moustache

 - Histogramme

Conclusion

Pourquoi résumer une variable ?

Objectif

En pratique, on résume une variable pour :

- ▶ **comparer** des groupes (ex : deux classes, deux conditions)
- ▶ **décrire** une population (valeur “typique”, dispersion)
- ▶ **détecter** des valeurs aberrantes / erreurs de mesure
- ▶ **préparer** la suite : corrélation, régression, ACP

Idée clé

Un bon résumé = **un nombre** + **un graphique** + **une explication**

Rappels : espace vectoriel \mathbb{R}^n

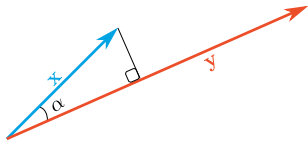
vecteur $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ transposé : $\mathbf{x}^\top = (x_1, \dots, x_n)$

norme $\|\mathbf{x}\|^2 = \sum_{i=1}^n x_i^2 = \mathbf{x}^\top \mathbf{x}$

produit scalaire $\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^n x_i y_i$

cosinus $\cos(\alpha) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$

projection $P_{\mathbf{y}}(\mathbf{x}) = \underbrace{\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{y}\| \|\mathbf{y}\|}}_{\mathbf{y}^\top \mathbf{y}} \mathbf{y} = \|\mathbf{x}\| \cos \alpha \frac{\mathbf{y}}{\|\mathbf{y}\|}$



Quel est le meilleur résumé d'une variable quantitative ?

► $S_5 = \{1, 5, 2, 4, 3\}$ $\mathbf{x} = (1, 5, 2, 4, 3)^\top \in \mathbb{R}^5$

Quelle est la valeur c qui **résume** au mieux les observations ?

Nous pouvons poser un principe variationnel :

$$\min_{c \in \mathbb{R}} J(c) \quad \text{avec :}$$

$$J(c) = \sum_{i=1}^n (x_i - c)^2 = \|\mathbf{x} - c \mathbf{1}\|^2$$

$$\Leftrightarrow \min_{c \in \mathbb{R}} \underbrace{(1 - c)^2 + (5 - c)^2 + (2 - c)^2 + (4 - c)^2 + (3 - c)^2}_{J_5(c) = 5c^2 - 30c + 55}$$

Quel est le meilleur résumé d'une variable quantitative ?

► $S_5 = \{1, 5, 2, 4, 3\}$ $\mathbf{x} = (1, 5, 2, 4, 3)^\top \in \mathbb{R}^5$

Quelle est la valeur c qui **résume** au mieux les observations ?

Nous pouvons poser un principe variationnel :

$$\min_{c \in \mathbb{R}} J(c) \quad \text{avec :}$$

$$J(c) = \sum_{i=1}^n (x_i - c)^2 = \|\mathbf{x} - c \mathbf{1}\|^2$$

$$\Leftrightarrow \min_{c \in \mathbb{R}} \underbrace{(1 - c)^2 + (5 - c)^2 + (2 - c)^2 + (4 - c)^2 + (3 - c)^2}_{J_5(c) = 5c^2 - 30c + 55}$$

► $S_6 = \{1, 5, 2, 4, 3, 100\}$ $J_6(c) = 6c^2 - 230c + 10055$

Quel est le meilleur résumé ?

Résumer S_n par $c \in \mathbb{R}$

$$S_n = \{x_1, x_2, \dots, x_i, \dots, x_n\}$$

Solution

$$\min_{c \in \mathbb{R}} J(c) \Leftrightarrow \frac{dJ(c)}{dc} = 0 \Leftrightarrow 2 \sum_{i=1}^n (x_i - c) = 0 \Leftrightarrow c = \frac{1}{n} \sum_{i=1}^n x_i$$

C'est la moyenne empirique : $c = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n \frac{1}{n} x_i = \sum_{i=1}^n \hat{f}_i x_i$

Définition : Moyenne

On appelle moyenne d'un échantillon x_1, x_2, \dots, x_n .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

(Dans le cas d'une variable qualitative ça n'a pas de sens.)

La Moyenne II

Propriété : si x_i sont regroupées en modalités m_i on a aussi $\bar{x} = \sum_{i=1}^r m_i f_i$ où m_i :

- ▶ est la valeur de la modalité m_i si la variable est discrète,
- ▶ est la moyenne des valeurs de la modalité,

$$m_i = [20 - 30] \quad \rightarrow \quad m_i = \frac{1}{n_i} \sum_{j \in m_i} x_j$$

Moyenne théorique

La moyenne **théorique** c'est l'espérance : $\mathbb{E}(X) = \lim_{n \rightarrow \infty} \bar{x}$

Propriété : $\mathbb{E}(X) = \int xP(x)dx$.

L'exemple de pile ou face illustre bien la différence entre moyenne théorique (espérance) et moyenne empirique.

Existe t'il autre chose que la moyenne pour parler d'un ensemble de valeurs ?

Surtout quand la moyenne dit des bêtises !

Moyenne : $\{1, 5, 2, 4, 3, 100\} = 115/6 = 19,17$

Médiane (le point milieu)

Définition : Médiane

La médiane est la valeur \widehat{M} telle que :

$$\widehat{F}_X(\widehat{M}) = \widehat{\mathbb{P}}(X \leq \widehat{M}) = \frac{1}{2}$$

(si on tire une valeur au hasard dans l'échantillon, on a autant de chance d'être au dessous que au dessus.)

Attention : deux cas pratiques

- ▶ **Données non regroupées** (souvent : variable continue mesurée, ou variable discrète sans effectifs > 1) :
 - on trie les observations ;
 - ▶ si n impair : valeur centrale,
 - ▶ si n pair : moyenne des deux valeurs centrales.
- ▶ **Données regroupées** (table de fréquences / classes) :
 - la médiane se lit via \widehat{F}_X (fréquences cumulées) et peut être **interpolée**.

Médiane (le point milieu)

x_i	n_i	f_i	F_i	\widehat{F}_X
1	1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$
2	1	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{12}$
3	1	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{5}{12}$
4	1	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{7}{12}$
5	1	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{9}{12}$
100	1	$\frac{1}{6}$	1	$\frac{11}{12}$

$$\widehat{F}_X(3) = 0,42$$

$$\widehat{F}_X(4) = 0,58$$

$$\widehat{M} = 3,5$$

Ici, $\widehat{M} = 3,5$ n'est pas une valeur observée : c'est normal.

La médiane est une **position** : elle peut **ne pas être** une valeur observée.

Médiane

Médiane **Théorique**

$$\min_{c \in \mathbb{R}} J_t(c) \quad \text{avec} \quad J_t(c) = \mathbb{E} (|X - c|)$$

Médiane **Empirique**

on remplace l'espérance par la moyenne :

$$\min_{c \in \mathbb{R}} J(c) \quad \text{avec} \quad J(c) = \frac{1}{n} \sum_{i=1}^n |x_i - c|$$

$$\frac{\partial J(c)}{\partial c} = 0 \Leftrightarrow \sum_{i=1}^n \text{signe}(x_i - c) = 0$$

Remarque : La médiane est plus robuste aux valeurs extrêmes

Définition : Mode

Le mode \widehat{M}_a d'un échantillon est égal à :

$$\arg \max_{x \in \Omega} \{\Pr(x)\}$$

- ▶ Dans le cas discret, c'est la valeur la plus fréquente :

$$\widehat{M}_a = m_i | i = \arg \max_{i=1, \dots, r} \hat{f}_i$$

- ▶ Dans le cas continu, c'est le **pic** de la distribution d'une variable continue.

Mode II

Remarques :

- ▶ La définition du mode n'exige pas de la variable qu'elle soit quantitative.
- ▶ Variable continue : le mode dépend de la méthode (bins de l'histogramme, largeur de classe, estimation par noyau). En continu : on parle souvent plutôt de densité/pics que d'un mode "exact".
- ▶ Multimodal : il peut y avoir plusieurs modes (plusieurs "pics").
- ▶ Sensibilité : une légère modification des données peut déplacer le mode.

Résumé central

À ces objets empiriques, on peut associer des objets théoriques.

- ▶ moyenne : la moyenne empirique / l'espérance

$$\min_{c \in \mathbf{R}} \sum_{i=1}^n (x_i - c)^2; \quad c = \sum_{i=1}^n f_i x_i \quad \mathbb{E}(X) = \int x \mathbb{P}(x) dx$$

- ▶ médiane : les fréquences cumulées / la fonction de répartition

$$\min_{M \in \mathbf{R}} \frac{1}{n} \sum_{i=1}^n |x_i - M|; \quad \widehat{\mathbb{P}}(X \leq M) = \frac{1}{2} \quad \mathbb{P}(X \leq M) = \frac{1}{2}$$

- ▶ mode : les fréquences / les probabilités

$$\arg \max_{i \in \{1, \dots, n\}} f_i \quad \arg \max_{x \in \Omega} \mathbb{P}(x)$$

Résumé central

Quand calculer quel type d'indicateur ?

Type de variable	Exemple	Moyenne	Médiane	Mode
Qualitative multimodale	les départements...			
Qualitative bimodale	oui/non (0/1)			
Ordinale	j'aime...			
Quantitative discrète	nombre...			
Quantitative continue	temps...			

Résumé central

Quand calculer quel type d'indicateur ?

Type de variable	Exemple	Moyenne	Médiane	Mode
Qualitative multimodale	les départements...	–	–	OK
Qualitative bimodale	oui/non (0/1)	OK	–	ok
Ordinale	j'aime...	ok	OK	ok
Quantitative discrète	nombre...	OK	OK	OK
Quantitative continue	temps...	OK	OK	–

Et une fois que l'on dispose d'une
valeur centrale ?

Surtout quand c'est la même !

Moyenne : $\{1, 5, 2, 4, 3\} = 15/5 = 3$

Moyenne : $\{3.1, 3, 2.9, 2.8, 3.2\} = 3$

Calcul d'un paramètre de dispersion

On considère l'échantillon $S_n = \{x_1, x_2, \dots, x_i, \dots, x_n\}$ de moyenne $\bar{x} = c$

Définition : Variance

C'est la moyenne des carrés des écarts à la moyenne :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

▶ Écart : $e_i = (x_i - \bar{x}) \Rightarrow$ Écart carré : $e_i^2 = (x_i - \bar{x})^2$

▶ Variance : $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n f_i e_i^2$

Note pour les calculs :

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n} \sum_{i=1}^n (x_i^2) - 2\left[\frac{1}{n} \sum_{i=1}^n (x_i)\right]\bar{x} + \bar{x}^2 = \frac{1}{n} \sum_{i=1}^n (x_i)^2 - \bar{x}^2\end{aligned}$$

Usage de la variance : donner une idée du domaine des observations I

Distance en
nb d'écart type

$$\mathbb{P} (|X - \bar{x}| < k\sigma) = 1 - \alpha$$

Écart d'une obs
à la moyenne

Probabilité

Usage de la variance : donner une idée du domaine des observations II

Cas Gaussien

pour un grand échantillon ($n \geq 100$) :

$$\mathbb{P}(|X - \bar{x}| \leq u_{\alpha/2} \sigma) = 1 - \alpha$$

avec $u_{\alpha/2}$ pris dans la table de la loi normale.

- ▶ Si on fixe $\alpha = 0,05$ (5%) $\implies k = 1,96$
 \implies 95% des observations $\in [\bar{x} - 1,96\hat{\sigma}, \bar{x} + 1,96\hat{\sigma}]$
- ▶ Si maintenant on fixe $k = 3$:
 - ▶ $\mathbb{P}(X - \bar{x} < 3\sigma) = 0.99865$
 - ▶ $\implies \mathbb{P}(X - \bar{x} < -3\sigma) = 1 - 0.99865 = 0.00135$
 - ▶ \implies 99,73% des observations $\in [\bar{x} - 3\hat{\sigma}, \bar{x} + 3\hat{\sigma}]$

Les quantiles I

Définition : Quantiles

On appelle quantiles (fractiles) à l'ordre p , $\forall p \in [0, 1]$, $\widehat{\Phi}_p$

$$\widehat{\mathbb{P}}(X \leq \widehat{\Phi}_p) = p$$

ou de manière équivalente : $\widehat{\Phi}_p$ telle que $\widehat{F}_X(\widehat{\Phi}_p) = p$

Les Quartiles

- ▶ $\widehat{\Phi}_{\frac{1}{4}} = \widehat{Q}_1$, telle que $\widehat{F}(\widehat{Q}_1) = \frac{1}{4}$,
- ▶ $\widehat{\Phi}_{\frac{1}{2}} = \widehat{Q}_2 = \widehat{M}$, telle que $\widehat{F}(\widehat{M}) = \frac{1}{2}$,
- ▶ $\widehat{\Phi}_{\frac{3}{4}} = \widehat{Q}_3$, telle que $\widehat{F}(\widehat{Q}_2) = \frac{3}{4}$.

Définition : Distance inter quartile (DIQ)

$$DIQ = \hat{Q}_3 - \hat{Q}_1$$

Résumé robuste

Considérons l'échantillon suivant : 4, 1, 6, 2, 72, 4, 6, 5, 1, 3, 7

Statistique	Échantillon complet	Sans la valeur 72
Moyenne	10.09	3.9
Médiane	4	4
Variance	425.69	4.54
Distance interquartile	3.37	4

Plan

Description mono variable

- La démarche descriptive

- Tableau de données

- Résumé (statistique) des données

 - Tendance centrale

 - Dispersion

 - Autres moments

 - Résumé robuste

- Résumé graphique des données

 - Cas des variables qualitatives

 - Boite à moustache

 - Histogramme

Conclusion

Pourquoi un résumé graphique ?

Un chiffre ne suffit pas

Deux distributions peuvent avoir la même moyenne (et même variance), mais une forme très différente.

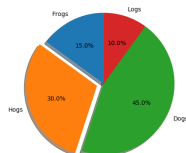
- ▶ Histogramme : forme globale, asymétrie, multimodalité
- ▶ Boîte à moustaches : médiane, dispersion, valeurs aberrantes
- ▶ Camembert : répartition des modalités (var. qualitative)

Bon conseil

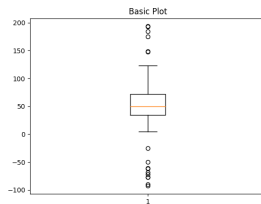
Regarder un graphique avant de conclure sur un indicateur.

Résumé graphique des données

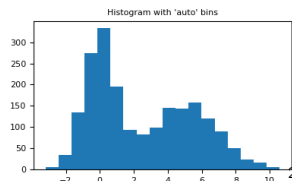
- ▶ Qualitatives :
 - ▶ Camemberts



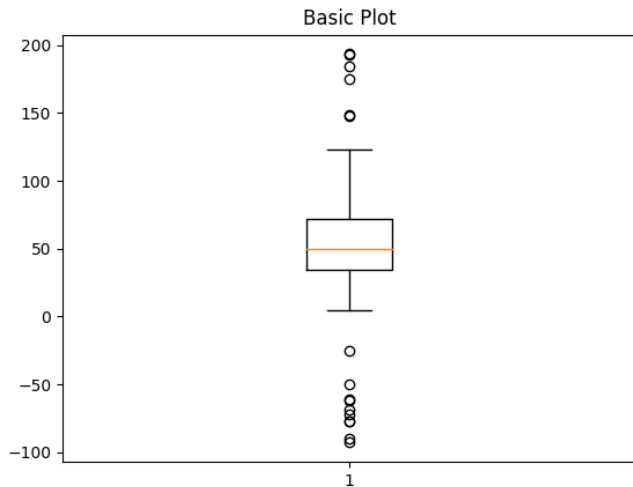
- ▶ Quantitatives :
 - ▶ Boîte à moustaches



- ▶ Histogramme
(discret ou continu après
discrétisation)



Variables quantitatives : boîte à moustache



Boite à moustache

Boite

- ▶ Ligne : médiane
- ▶ Extrémités de la boîte : Quartiles

Épures (whiskers)

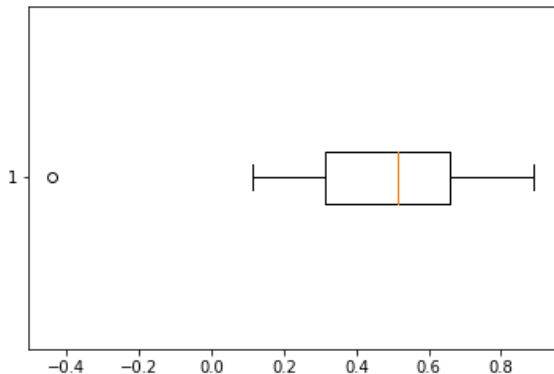
- ▶ Par défaut :

$$[\hat{Q}_1 - \frac{3}{2}DIQ; \hat{Q}_3 + \frac{3}{2}DIQ].$$

- ▶ Réglable : RTFM : https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.boxplot.html
- ▶ Les points hors épure sont représentés (outliers)

Boite à moustache

0.5300
0.2841
0.4361
0.1869
0.8340
0.1406
0.6683
0.8088
0.3240
0.3618
0.3096
0.5754
0.2800
-0.720
0.6135
0.5610
0.4228
0.1138
0.8681
0.8665
0.5023
0.6334
0.8877

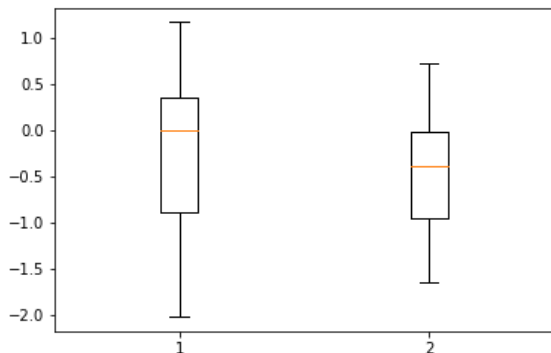


Exemple d'utilisation des boîtes à moustache

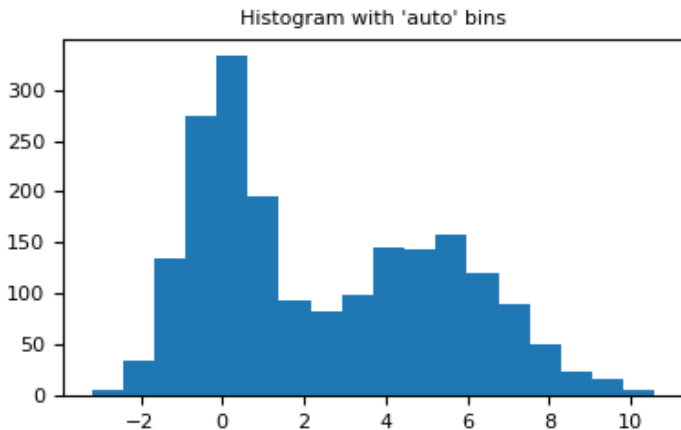
-1.1754	-0.6376
-0.3977	-0.3052
-1.9512	0.5161
0.9047	0.3681
0.6822	0.0715
-2.0183	-0.3123
-0.4543	0.7366
0.0421	-0.3635
-0.4582	-0.9618
0.2641	-0.2175
0.2920	0.0235
0.3868	-0.7281
0.5540	-1.3258
0.9437	-0.1650
0.0598	-0.1500
0.5747	0.5769
-1.4419	0.2333
-0.2355	-0.9544
-1.0214	-0.4148
0.1398	-1.0734
0.0816	-1.0024
-1.2582	-0.7818
-1.4685	-1.0750
-0.0329	-1.6426
-1.6925	-0.9326
-0.4757	-0.8195

Pour comparer visuellement 2 variables :

$$c_1 = -0.13 \quad c_2 = -0.55 \quad \hat{\sigma}_1 = 1.02 \quad \hat{\sigma}_2 = 0.67$$



Variables discrètes : Histogramme



- ▶ Abscisse : intervalle de valeurs
- ▶ Ordonnée : hauteur
- ▶ Surface : nombre d'individu ou fréquence

Histogramme : discrétisation (bins)

- ▶ **Variable discrète** : on compte directement les occurrences de chaque modalité.
- ▶ **Variable continue** : on **découpe** l'axe en p classes (bins).

Bins (classes)

On choisit des bornes $b_0 < b_1 < \dots < b_p$. Chaque barre correspond à l'intervalle $[b_{i-1}, b_i[$.

Hauteur / surface

Soit s_i le **nombre d'observations** dans $[b_{i-1}, b_i[$ et $\Delta_i = b_i - b_{i-1}$.

$$h_i = \frac{s_i}{\Delta_i} \quad \Rightarrow \quad \text{surface} = h_i \Delta_i = s_i$$

(**Variante** : si on normalise en fréquences, la surface totale vaut 1.)

Histogramme : choix du nombre de classes

Règle simple (pratique)

- ▶ On prend en général $p \approx \sqrt{n}$ (ordre de grandeur).
- ▶ Puis on **ajuste visuellement** :
 - ▶ trop petit p : histogramme “trop lissé” (perte de structure),
 - ▶ trop grand p : histogramme “trop bruité”.

Important

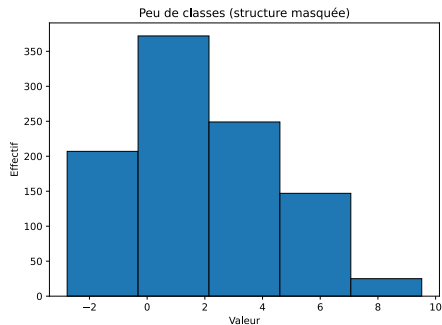
Si les classes n'ont pas la même largeur, il faut raisonner avec la **hauteur**

$h_i = \frac{s_i}{\Delta_i}$ (densité), pas seulement avec s_i .

Histogramme : influence du choix des classes

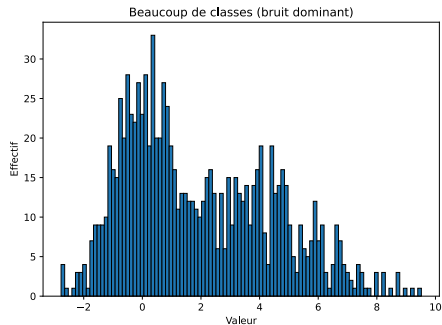
Peu de classes

(structure masquée)



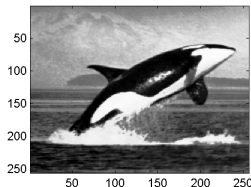
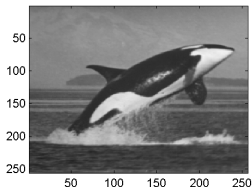
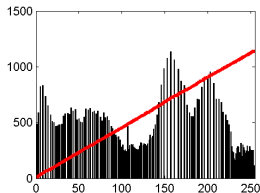
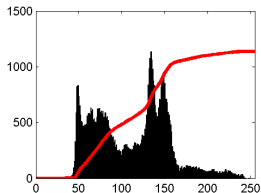
Beaucoup de classes

(bruit dominant)



Histogramme et traitement d'images

Redressement (où égalisation) d'histogramme¹ : $b_n(i) = \hat{F}^{-1} \left(\frac{i}{p} \right)$



1. http://en.wikipedia.org/wiki/Histogram_equalization

Histogramme et traitement d'images

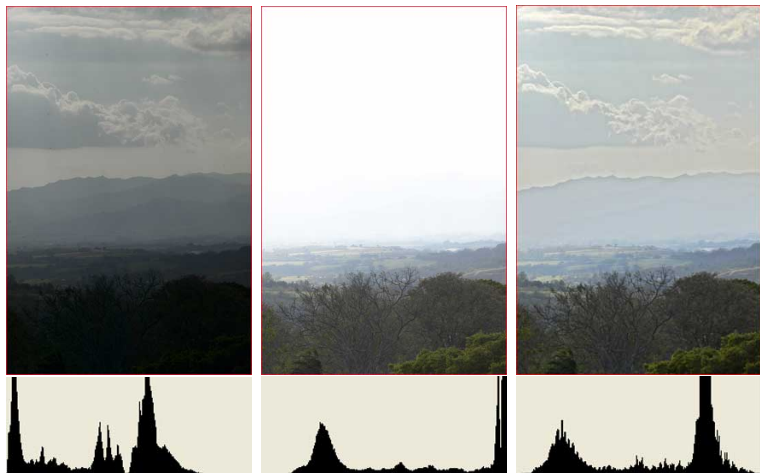


Figure – Exemple d'histogrammes <http://www.llvj.com/tutorials/understanding-series/understanding-histograms.shtml>.

Plan

Description mono variable

- La démarche descriptive

- Tableau de données

- Résumé (statistique) des données

 - Tendance centrale

 - Dispersion

 - Autres moments

 - Résumé robuste

- Résumé graphique des données

 - Cas des variables qualitatives

 - Boite à moustache

 - Histogramme

Conclusion

Empirique vs. Théorique

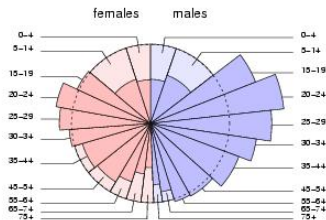
Cas d'une variable X discrète :

Nom Empirique	Empirique	Théorique	Nom Théorique
Fréquence	$\hat{f}_i = \frac{n_i}{n}$	$\mathbb{P}(X = m_i)$	Probabilité
Fonction de répartition	$\hat{F}_i = \frac{N_i}{N}$	$\mathbb{P}(X < a_i) = F(m_i)$	Fonction de répartition
Moyenne	$\bar{x} = \sum_{i=1}^n f_i x_i$	$\mathbb{E}(X) = \sum_{i=1}^n \mathbb{P}(m_i) m_i$	Espérance
Médiane	$\hat{F}(\hat{M}) = \frac{1}{2}$	$\mathbb{P}(X < M) = \frac{1}{2}$	Médiane
Variance	$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$	$\sum_{i=1}^n \mathbb{P}(m_i) (m_i - \mathbb{E}[X])^2$	Variance

Cas d'une variable X continue : l'empirique ne change pas (avec $n_i = 1$)

Conclusion

- ▶ Résumé quantitatif
 - ▶ ordre 1 : centrage
 - ▶ ordre 2 : dispersion
 - ▶ ordres supérieurs : asymétrie, aplatissement...
- ▶ Résumé qualitatif (graphique)
 - ▶ qualitative : camembert
 - ▶ quantitative (continue) : boîte à moustache.
 - ▶ quantitative discrète : histogramme
- ▶ Méthode exploratrice
 - ▶ pour une exploration interactive des données

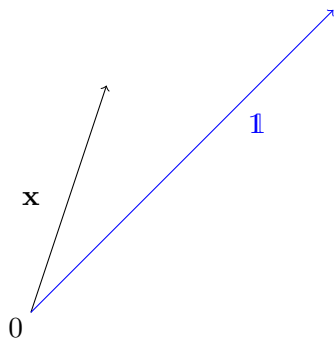


Interprétation géométrique de la moyenne

$$\begin{aligned} J(c) &= \sum_{i=1}^n (x_i - c)^2 \\ &= \|\mathbf{x} - c \mathbf{1}\|^2 \end{aligned}$$

► $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$

► $\mathbf{1} = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$



c^* , la moyenne est aussi la projection orthogonale du vecteur des observations sur le vecteur $\mathbf{1}$

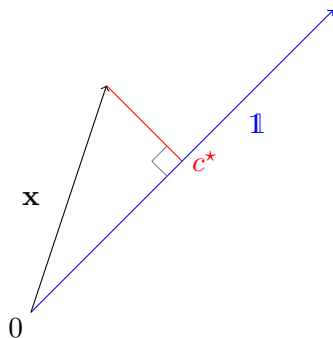
$$c^* = \frac{\mathbf{x}^\top \mathbf{1}}{\mathbf{1}^\top \mathbf{1}} = \frac{\sum_{i=1}^n x_i}{n}$$

Interprétation géométrique de la moyenne

$$\begin{aligned} J(c) &= \sum_{i=1}^n (x_i - c)^2 \\ &= \|\mathbf{x} - c \mathbf{1}\|^2 \end{aligned}$$

► $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$

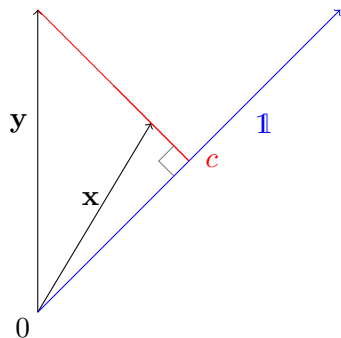
► $\mathbf{1} = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$



c^* , la moyenne est aussi la projection orthogonale du vecteur des observations sur le vecteur $\mathbf{1}$

$$c^* = \frac{\mathbf{x}^\top \mathbf{1}}{\mathbf{1}^\top \mathbf{1}} = \frac{\sum_{i=1}^n x_i}{n}$$

La moyenne des écarts à la moyenne : la variance



La variance $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{x} - c \mathbf{1}\|^2$$

Pour deux échantillons \mathbf{x} et \mathbf{y} de même moyenne, la variance traduit leur proximité avec le vecteur $\mathbf{1}$.

L'Écart type : $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

C'est la norme du vecteur $\mathbf{x} - c \mathbf{1}$ (divisée par \sqrt{n})

Et Pythagore nous dit que : $\|\mathbf{x}\|^2 = nc^2 + \|\mathbf{x} - c \mathbf{1}\|^2$